© 2019 IJSRST | Volume 6 | Issue 1 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology
DOI : https://doi.org/10.32628/IJSRST19615

Image Purification Technique for Myanmar OCR Applying Skew Angle Detection and Free Skew

Chit San Lwin[1], Wu Xiangqian[2]
[1,2]School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, P. R. China
[1]Department of Mathematics, Kyaing Tong University, Kyaing Tong City, Shan State, Myanmar
chitsanlwin.maths.mm@gmail.com[1] xqwu@hit.edu.cn[2]

ABSTRACT

Optical Character Recognition (OCR) is a technology widely adopted for automatic translation of hardcopy text to editable text. The language dependence of technology makes it far less developed for less popular languages like Myanmar language. Also, the uniqueness and complexity of the Myanmar text system such as touching and complex characters have continued to pose serious challenges to several OCR investigators. In this paper, we propose a new technique to development Myanmar OCR system. Our technique implement skew angle detection and free skew, noisy border correction, extra page elimination, line segmentation from scanned images of Myanmar text. Performance of the proposed method is tested with 430 documents comprising different printed and handwritten Myanmar text of various fonts, sizes, multi-column, tables, stamps or photos, background effects. Our method gives an accuracy of 100% for line segmentation and 99.92% for skew angle detection and free skew. The ability of our method to effectively implement global and local skew angle detection, free skew and line segmentation in different handwritten and digital text images of the Myanmar character set with high accuracies confirm the robustness of the technique, its reliability and its suitability for application in many other related languages.

Keywords: Myanmar Script, Document Image Analysis, Skew Angle Detection, Free Skew, Border Edge Discarding, Lines Segmentation

1. INTRODUCTION

In the last few decades, the need to reincarnate historic information held on papers into electronic forms for easy accessibility and sharing is increasing. The burden of maintaining a large volume of papers in an office space and the difficulty in transmitting the same within and outside the holding organization are driving factors behind this need. In most of the global and widely spoken languages like English, Chinese and Russian, this burden has been greatly reduced with solutions from OCR technology. However, the same cannot be said of fewer poplar languages like Myanmar language which has a comparatively unique and complex character set in the writing system. Hence there is a dare need for a robust and reliable OCR system to help preserve history and convert paper-based text to digital information.

OCR uses scanned images from automatic document feeders or photos taken with still cameras. In the process of scanning a text-based paper or taking a photo of it, errors either due to mechanical dysfunction in the capturing device or improper positioning of the camera, arise. In other situations, the captured image may contain some information that is non-textual or irrelevant to the textual information intended for extraction from the document. The purpose of OCR is to convert from image text into machine-readable text. However, when errors like this occur, it becomes difficult for the machine to correctly extract and read the textual information.

One common and dementing error experienced in these images is a skew angle. A situation that causes the text lines to tilt apart from the horizontal or vertical plane, thereby rendering an image that cannot seat parallel to the x-axis or y-axis. Skew angles are classified as global skew (also known as single/uniform skew where the text lines in the image take the same inclination angle as shown in Fig. 5(a-f, h)), local skew (where multiple text lines in the image shows various inclination angles as shown in Fig. 5(g)) and non-uniform text line skew (where the inclination forms a set of zigzag or curvilinear text lines in the image), [1]. When the images skewed in any of these forms, the accuracy of the OCR system becomes significantly dented. To get superior and high accuracy OCR outputs, images being passed to the OCR system have to be checked and corrected from any skews [2]. To overcome these, several methods for skew angle detection and free skew technique to improve the accuracy of OCR systems have been proposed, but these [3-9] do not discuss about it.

Generally, OCR systems comprise six steps leading to enhanced document images and useful information extraction [10]. These steps include image acquisition, preprocessing (noise removal, skew angle detection, and free skew), segmentation, feature extraction, recognition, and post-processing. In OCR systems, the effectiveness of each succeeding step depends on the accuracy of the preceding step; hence, the results of the skew angle detection and free skew depend on the proper pre-processing of the document image received as input. Before segmentation and feature extraction, a challenging task is to detect the skew angle and free skew of the document images, therefore getting this correctly is very important for preventing errors in the further steps of OCR image analysis [11].

The most popular methods for skew angle detection and free skew are Projection Profile Analysis, Fourier Transform, Nearest Neighbor Connectivity, Mathematical Morphology, Cross-Correlation, and Hough Transform among others [11]. Using insight from these methods, our implemented Myanmar OCR system is applied to correct and accurately extract text from document images having multiple columns, tables, photos, stamps, logos, background effects, global and local skews, flow-charts, border, and extra pages.

The remainder of this work is arranged as follows. In Section 2, we present the related OCR solutions reported in the literature and give a brief overview of the Myanmar text system in Section 3. This is followed by a description of our methodology in Section 4 and the results and discussion of the method are in Section 5. We conclude in Section 6 with our closing remarks.

2. LITERATURE SURVEY

Skew angle detection and free skew are crucial tasks in the preprocessing step of OCR systems. This is because their outcome directly affects the results of the OCR system. Feeding an OCR system with image files containing skew reduces the text extraction and recognition accuracy of OCR system. To this end, many researchers have proposed techniques for skew angle detection and free skew for improving OCR accuracy.

Papandreou et al. [12] presented a novel skew detection technique based on vertical projections. This system attached bounding box minimization technique on languages having vertical stokes like Latin, German and French characters by rotating the vertical range of angles after finding the bounding box area. To the best of our knowledge, their technique will fail on languages, which own circle characters like Myanmar script [9, 13]. Furthermore, Papandreou et al. [3] extended the work [12] by presenting a new skew detection technique that fuses horizontal projection profiles for printed document images based on a combination of enhanced projection profiles (CEPP). They rotated

bounding boxes with respect to the horizontal as well as the vertical range of angles. They chose the minimum bounding box as a skew angle by comparing every angle after calculating each angle and used a coarse-to-fine technique (CEPPc-f) to accelerate their algorithm. The methods [12] and [3] work well on images with noise and warp. However, their techniques will fail on images files having multiple skews.

Mohammed et al. [14] proposed global skew detection and correction using morphological and statistical methods. They firstly defined structuring element (SE) by using morphological dilations. They then pull out the longest connected components and seek the global skew by using statistical analysis on it. They tested the performance of their method on printed and handwritten images of English, Devanagari and Arabic texts. They however did not cover multiple skews in their submission. Soora et al. [15] proposed a novel local skew correction and segmentation approach for printed multilingual Indian documents by using a similar technique to [14]. They achieved 97% and 94% accuracy on 330 documents having English, Devanagari, Marathi, Telugu and Kannada scripts for lines and words segmentation respectively. Their presentation however did not include image files with photos, graphics, logos or stamps, hence we suppose their method cannot efficiently handle text extraction from files having these extra features.

Singh et al. [16] exhibited the improved skew detection and correction approach using Discrete Fourier Transform (DFT) and angle of elevation theory. Initially, they calculated the Fourier spectrum in each block by using fast Fourier transform (FFT) after splitting the input image into 4-blocks with corners. Next, they designated every skew angle by assessing the highest frequency of FT in each block. Finally, they subtracted the skew angle for the whole image by taking the mean of 4-block skew. Due to being the ability of DFT transmitted from FFT and FFT inverse, their algorithm is fast. They tested their method on 120 images having different languages (English, Hindi, and Punjabi) with pictures. Similarly, Watts et al. [4] presented a performance evaluation to improve skew detection and correction using FFT and median filtering. To reduce computing time in their algorithm, they employed the DCT technique. Their technique produced a 99% accuracy with 25 skewed images. However, the proposed methods [16] and [4] ignore both handwritten image files and printed images having multiple orientations.

Boukharouba et al. [17] proposed an algorithm for skew correction and baseline detection based on randomized Hough Transform (RHT). The study utilized 400 text images with various font types and sizes as a dataset. The method detected a lower baseline of text lines after applying edge on the grayscale text images. The skew angle of each grayscale image was then determined by RHT. Their method, as reported, had a 100% accuracy for skew angle detection and correction, while the baseline detection technique produced an accuracy of 95%. However, their dataset did not include text images containing figures, tables, logos and/or stamps.

Similarly, the authors [18] applied a straight-line Hough Transform technique for skew angle detection and correction for the Telugu language. Using Morphological Skeleton and Progressive Probabilistic Hough Transform (PPHT), Omar et al. [5] also proposed a skew angle detection and correction technique for historical scanned documents. They basically applied the Morphological Skeleton technique for iterative thinning of the binary image. Whereas, PPHT was employed for line detection and estimation of global skew in each image. Although high accuracy was achieved in skew angle detection and free skew on images having different languages with figures and

tables, the submission did not cover multiple skews and handwritten document images.

Furthermore, the authors [6] proposed a technique for skew angle detection and correction by applying the rectangle method. Their technique considering the image as a matrix of pixels scanned in each de-noised image to determine the minRow, maxRow, minColumn and maxColumn in order to create a rectangle representative of the respective image. They detected skew angle by using bisection numerical method. Their results show that the method is fast and reliable in relation to skew angle detection and correction. Likewise, Jundale et al. [7] proposed skew detection on scanned documents written in the Devanagari language. Their technique used pixels of an axis-parallel rectangle and linear regression for skew angle detection of word/line. An accuracy of 91% was obtained for words and lines in handwritten Devanagari document images. The methods [6] and [7] can deal on images with photos and tables. However, their algorithms will fail on images having multiple skews.

Using the piece-wise painting algorithm (PPA), Alaei et al. proposed a skew angle detection technique [8]. The technique used PPA to determine and choose specific regions from horizontal and vertical painted images then applied linear regression and line drawing methods to pull out two fit lines. Following, they calculated skew angle after taking the best-fit line by voting approach on fit lines and applied iteration to achieve higher accuracy. Their technique was tested with 743 text images including English, Chinese and Japanese language texts with tables, figures, etc. The images were mostly handwritten; having figures shapes and tables alongside the document text. Although they got 93.2% accuracy on it, their method cannot detect skew angle in images with multiple skews.

In addition, Hasan et al. [19] proposed a technique for Bangla printed text by document decomposition. They traced edges in images by applying the Canny algorithm. Horizontal and vertical areas were determined for each grayscale image, after which image pixels were split into textual and non-textual zones. Using over 300 text images with photos, headlines, sub-headlines and columns; their technique produced an accuracy of 97.02% for document decomposition. Lastly, the submissions [1, 11, 20-22], surveyed skew angle detection techniques and algorithms, highlighting the weakness and strengths of different proposed skew angle detection and correction algorithms. From the above, we observe that most of the proposed methods simply rotate the skewed image against the direction of the skew angle. This technique is limited arrangement. Nevertheless, our proposed method intends to cover it in the Myanmar language.

3. BACKGROUND OF MYANMAR SCRIPT

According to United Nation estimates [23], Myanmar has a population of about 54 million people on 12 April 2018, with 135 ethnic groups. There are approximately a hundred languages spoken in Myanmar [24]. The official Myanmar language (known as Burmese) is spoken by 33 million people as a first language. Moreover, about 10 million people, particularly ethnic minorities in Myanmar and those in neighboring countries, use Burmese as their second language [25]. The Myanmar language is a member of the Lolo-Burmese grouping of the Sino-Tibetan language family. The Burmese alphabets ultimately descended from a Brahmic script, either Kadamba or Pallava of South India and more immediately an adaptation of Old Mon or Pyu script. Burmese is a tonal, pitch-register and syllable-timed language that is largely monosyllabic and analytic, with a subject-object-verb word order [25]. The Myanmar alphabets

are also being used for the liturgical languages of Pali and Sanskrit.

Burmese is a diglossic language with two distinguishable registers (diglossic varieties) being Literary High form (formal and written used in literature, newspapers, radio broadcasts and formal speeches) and Spoken Low form (informal and spoken used in daily conversation, television, comics and literature (informal writing)). The literary form of Burmese retains archaic and conservatively grammatical structures and modifiers (including particles, markers, and pronouns) no longer used in the colloquial form. In most cases, the corresponding grammatical markers in the literary and spoken forms are totally unrelated to each other. Nowadays, television news broadcasts, comics, and commercial publications are using the spoken form or a combination of the spoken and simpler, less ornate formal forms [25].

Myanmar script has 34 Basic Consonants including "ဥ", 12 Independent Letters including "ဩ" and "ဧ", 12 Vowels, 4 Medial or Semi-vowels, 2 Primary Break Characters (as a comma and a full stop in English), 10 digits, 11 Extensions of Pali and Sanskrit and 75 Extension Characters of some ethnics (such as Mon, Shan, Karen, Kayah and Palaung in Myanmar). Range of Myanmar Unicode is U+100 to U+109F (160 code points) for the Unicode Standard, Version 10.0, [26] shown in Fig. 1. Myanmar script does not have lower and upper cases as in English. The direction of the writing system is from left to right in horizontally and requires no spaces between words, although modern writing usually contains spaces after each clause to enhance readability.

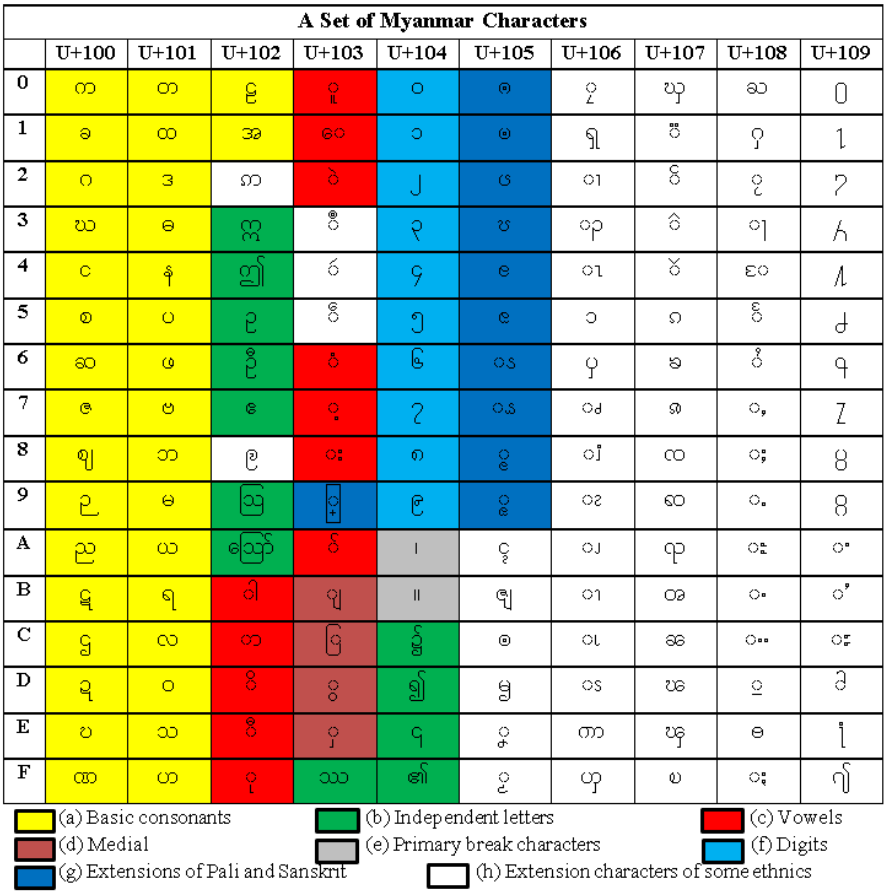


Figure 1. Myanmar language characteristics.

4. PROPOSED METHODOLOGY

The methodology for skew angle detection and free skew checking and correction system described in this study is multipartite. Tasks include Image Acquisition, Pre-processing, Skew Angle Detection, Free Skew, Discarding of Borders and Extra Pages and Checking System of them. Fig. 2 is a description of the task flow.

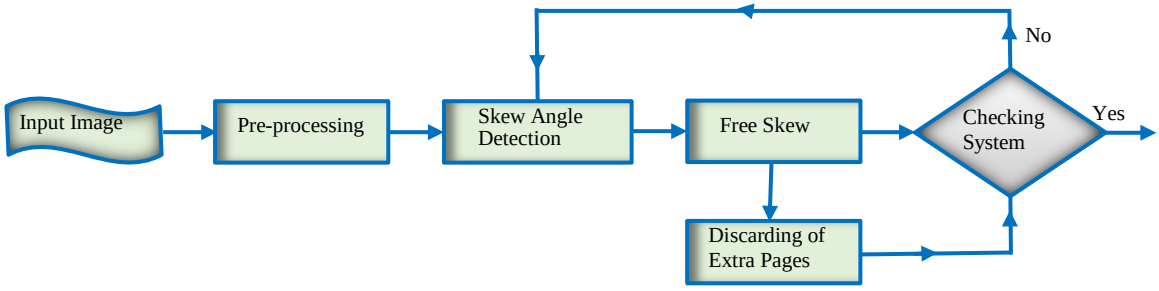


Figure 2. Depict of our proposed methodology.

4.1 IMAGE ACQUISITION

Though image acquisition for this work is basic, the accuracy of the results depends strongly on the quality of the images supplied as input [14]; hence, this task is very essential. To this end, we acquired 430 images comprising various font types and sizes, text regions more than one column, graphs, tables, and photos or images from Myanmar daily newspapers and web-based publicly accessible social platforms. In addition, handwritten Myanmar text of different lengths was collected from 15 writers. All hardcopy text files were collected by using the flatbed high-resolution scanner or phone camera, while publicly accessed image files corresponded to .jpg, .png and .bmp in extensions with 300-dpi and 600-dpi pixel distribution.

4.2 PREPROCESSING

An initial and necessary step of any recognition system is preprocessing. The main task in this step is noise removal and quality enhancement of all input images. Also, pre-processing helps regularize input image size to a specified standard for easy and error-free utilization in the system. In our work, pre-processing included binarization and noise removal among others.

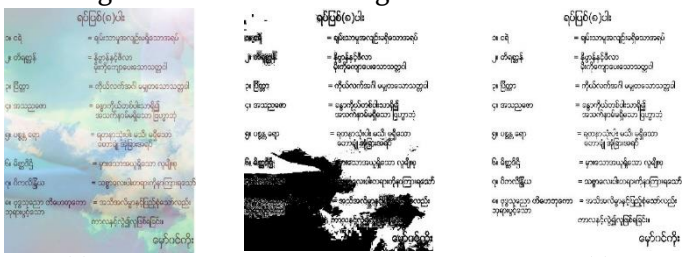
4.2.1 BINARIZATION

Whether the input image to an OCR system is colored or grayscale, it has to be converted to its binary equivalent in order to discard every hue and enhance its brightness. In other words, each input image is split into the foreground (black) and background (white) [1]. We call this process as digitization or bi-level or two-level image preprocessing. It reduces the overall computational time and noises in images. As such, a robust binarization technique is required for an OCR system to produce efficient results.

There are several binarization algorithms, but most existing works [8, 13, 27] have been using the Otsu [28] method. However, this method cannot give a satisfactory result when the input image had a shadow region or shadow background and lighting effects. But in [5] an effective binarization technique (Nick method) is proposed using the following formula

$$T = M + F\sqrt{\frac{\sum_{i=1}^{N}\left(p_i^{\,2} - M^{2}\right)}{N}},$$

where F is the factor $(-0.2 \le F \le -0.1)$, p_i = pixel value of the grayscale image, N = the number of pixels in the window and M = mean gray value of these N pixels. The output result of this method surpasses the limitations of binarization with the Otsu algorithm. In our work, we adopt the Otsu and Nick methods and the results of our technique compared to the result of the Otsu algorithm is shown in Fig. 3.



(a) (b) (c)

Figure 3. Comparison results of Otsu Thresholding and our methods; (a) Original image, (b) Otsu method, and (c) Ours.

4.2.2 REMOVAL OF NOISE

Noise emerges on acquired images due to printer or scanner errors, poor paper quality or ink, and excess dust on papers or due to ancient document formatting style. Such noises in input images result in blurs and isolated pixels among other effects that limit the accuracy of results produced by the OCR system. To effectively manage such noises; several algorithms including Chebyshev filter, Elliptic (Cauer) filter, Butterworth filter, Bessel filter and so on have been proposed. However, we opted for the Gaussian filter [29-31] to discard emergent noises in input images. The Gaussian filter is described by the following equation.

$$g(x,y) = \frac{1}{2\pi\sigma^2} \cdot e^{-\frac{x^2+y^2}{2\sigma^2}}, \quad (1)$$

where σ is the standard deviation of the Gaussian distribution and x is the distance from the origin in the horizontal axis and y is the distance from the origin on the vertical axis of pixels contained in the image.

4.3 SKEW ANGLE DETECTION

After the binarization step, skew angle detection is performed for Myanmar text document images. There exists a skew angle in a document image when the text lines tilt apart from the horizontal or vertical axis. This problem is a major challenge for researchers in document analysis, especially for the Myanmar language, it is an unsolved problem. Skew angle detection, which the most of researchers used in his pre-processing stage to get image points or image pixels is the main problem in all document analysis and recognition system because free skew directly affects the future steps of document analysis and OCR system.

Several researchers have proposed diverse methods for skew angle detection and free skew correction, [3-8, 15, 39]; but in this study, we adopted Hough Transform proposed by Paul Hough in 1962 because of its comparative robustness and high accuracy even though it increases computational time. Besides, it has been used in vehicle license plate recognition system to remove skew angle [32]. The tradeoff therefore is between achieving a low accuracy with fast processing or a high accuracy with longer computation time. Our proposed algorithm for skew angle detection and free skew of Myanmar text document images is very adaptive and takes accuracy over speed.

For any text line in the image file that has one variable, we can generally express it as a line equation form;

$$ax = b \text{ or } ay = b, \quad (2)$$

where a and b are constants but $a \neq 0$. In addition, it has two variables, the equation of the line is given in the form;

$$y = mx + b, \quad (3)$$

where b is the point of that line passing through the y-axis, also known as the y-intercept and m is either the slope or gradient of that line and it is given in the form;

$$m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{\Delta y}{\Delta x} = \frac{rise}{run}, \quad (4)$$

where (x_1, y_1) and (x_2, y_2) are two points on the line such that $x_1 \neq x_2$, [33, 34]. It is also known as the slope-intercept form, which described a straight line in the Hough Transform method [35]. In trigonometry, we can also write the slope of the line as a next version like $m = \tan(\theta)$ with its inclination angle θ, [34]. As mention in Eq. (3), in Cartesian space, we can express in Fig. 4(a) as well as in the Hough space can depict as Fig. 4(b). Moreover, it can also express as Eq. (5) in polar,

$$r = x\cos\theta + y\sin\theta, \quad (5)$$

where r is the orthogonal distance from the origin to the point laying on the line and θ is an angle between x-axis and r joining original and a point on the line.

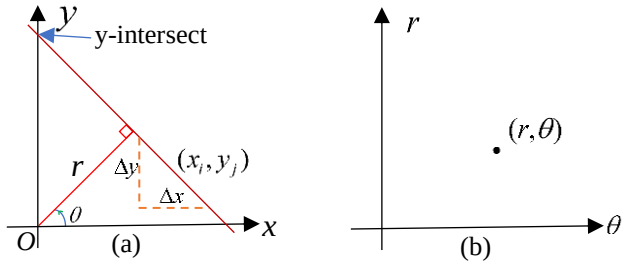


Figure 4. Depiction of Cartesian and Hough spaces; (a) Cartesian space, and (b) Hough space.

The line in the image space becomes a point in the Hough space as shown in Fig. 4. The linear Hough transform algorithm applies two-dimensional array (accumulator array) to find the existence of a line in image space [35] where each row and column equal to r and θ values. The details of our method are described in Alg. 1. After finding the slope in the Myanmar text document images using linear regression, we then calculate for the skew angle by applying Eq. (6),

$$\theta = \tan^{-1}(m) \times \frac{180}{\pi}. \quad (6)$$

The second part of Eq. (6) aims to change the degree. Fig. 5 is been described as sample images having a various skew of input Myanmar text documents.

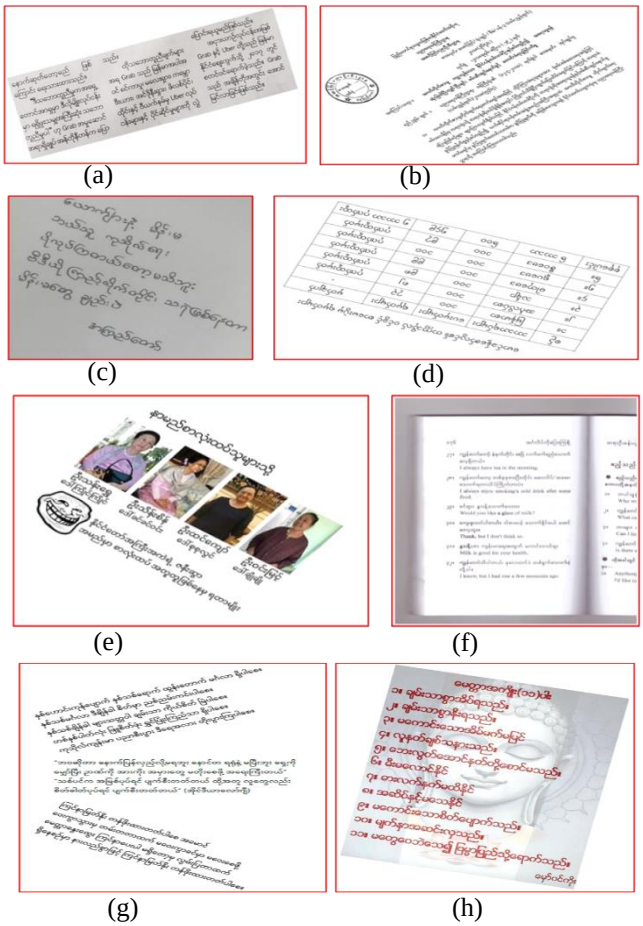
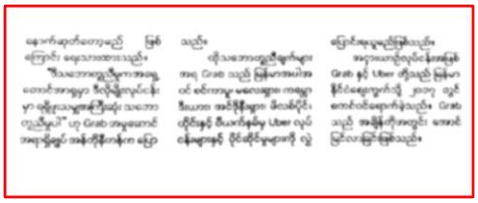


Figure 5. Myanmar text document images with various skew angle; (a) 3-columns image in the newspaper, (b) Scanned image with the stamp from prohibition law, (c) Handwritten image of Myanmar text, (d) Inclination image with table, (e) Tilted text image with photos, (f) Scanned image from a thick book, (g) Myanmar text image with global skew, and (h) Image with background effects.

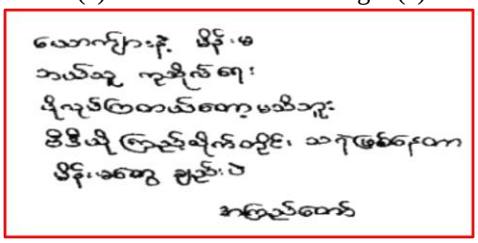
4.4 FREE SKEW

After determining the skew angle in the input text images, the files are rotated to get free skew. This is done to ensure the text image files aligns properly to the horizontal or vertical axis, i.e., the tilted angle (θ) of the text images should be approximately zero (0) or perpendicular to the axes. In most existing works in literature, file cropping method has been used to effect free skew. This method is liable to cause information

loss, as a vital part of the text in the images might be cropped off in the process; leading to reduced accuracy in results. Our proposed method as described in Alg. 1 for skew angle detection and free skew execution overcomes the highlighted limitations of the cropping method. Results of our proposed technique for free skew are presented in Fig. 6, while Fig. 7 compares our result with the result of cropping method.



(a) Free skew result of Fig. 5(a)

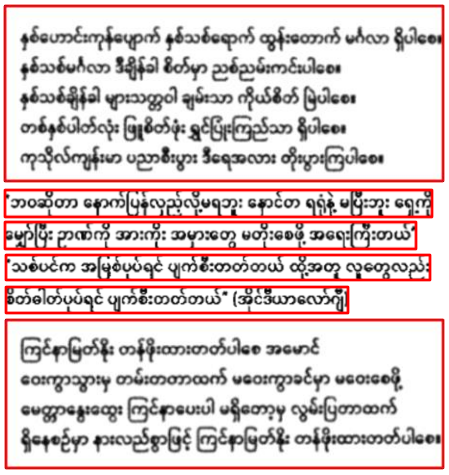


(c) Free skew result of Fig. 5(c)

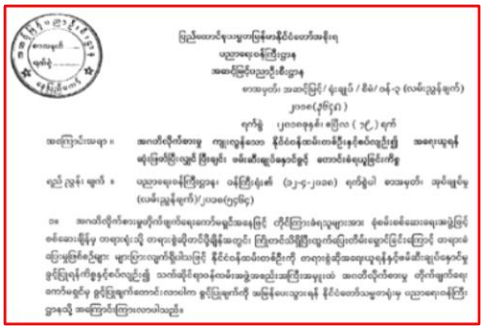
မောင်ခန့်စော်လွင်၏ တက္ကသိုလ် ဝင်ခွင့် စာမေးပွဲမှ ရမှတ်များ

စဉ်	ဘာသာရပ်များ	ပေးမှတ်များ	ရမှတ်များ	မှတ်ချက်
၁။	မြန်မာစာ	၁၀၀	၇၃	-
၂။	အင်္ဂလိပ်စာ	၁၀၀	၈၂	ဂုဏ်ထူးမှတ်
၃။	သခ္ခါ	၁၀၀	၉၈	ဂုဏ်ထူးမှတ်
၄။	ဓါတုဗေဒ	၁၀၀	၉၉	ဂုဏ်ထူးမှတ်
၅။	ရူပဗေဒ	၁၀၀	၁၀၀	ဂုဏ်ထူးမှတ်
၆။	ဇီဝဗေဒ	၁၀၀	၉၇	ဂုဏ်ထူးမှတ်
စုစုပေါင်း	၆ ဘာသာ	၆၀၀	၅၄၉	၅ ဘာသာ ဂုဏ်ထူး

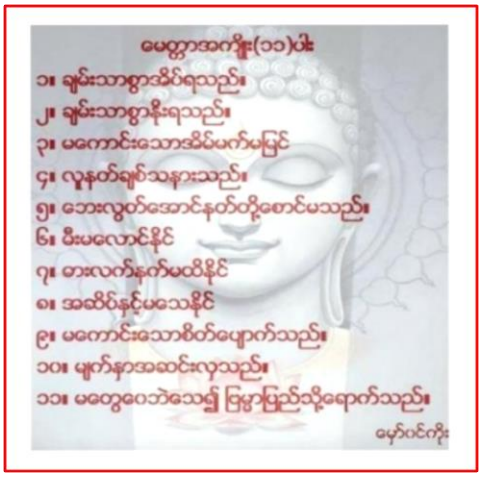
(d) Free skew result of Fig. 5(d)



(f) Free skew result of Fig. 5(g)



(b) Free skew result of Fig. 5(b)



(e) Free skew result of Fig. 5(h)

(g) Free skew result of Fig. 5(e)

Figure 6. Depiction of free skew results of Fig. 5 having various skew.

Algorithm 1

1. *Read Myanmar Text Document.*
2. *If RGB then*
   gImg ⟵ *convert to gray of the orig image*
   end.
3. *BPixels* ⟵ *gImg less than some standard thresholds.*
4. *[x, y]* ⟵ *find indexes in the BPixels.*
5. *cofs* ⟵ *use the polyfit function for the first degree of x and y.*
6. *xEnds* ⟵ *columns of the first row.*
7. *yEnds* ⟵ *use the polyval function to calculate the polynomial cofs at xEnds.*
8. $\theta \longleftarrow \tan^{-1}\left(\frac{yEnds(2)-yEnds(1)}{xEnds(2)-xEnds(1)}\right)$.
9. *ImgTurn* ⟵ *rotates BPixels according to against θ by bilinear method.*
10. *Img* ⟵ *save change ImgTurn into black on white image.*
11. *kk* ⟵ *do brightness color of Img image.*
12. *Save kk image.*

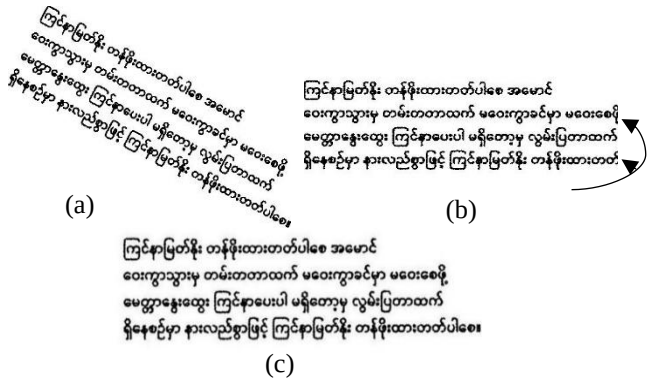


Figure 7. Comparison of cropping method and ours; (a) Input text image with inclination, (b) Loosed information by cropping method, and (c) Effectiveness of proposed method.

4.5 REMOVAL OF BORDERS AND EXTRA PAGES

When the thick books are scanned to get the document images, the resulting images may emerge with gray or black borders and extra pages among many other noisy effects. Due to mechanical errors from the scanner, document images can have above-mentioned noisy effects. Some example of these uselessness effects on text images is illustrated in Fig. 5(f) and Fig. 8(a) respectively. Without eliminating or properly managing these defects; determination of skew angle and free skew becomes difficult, and the accuracy of the results of the OCR system will be decreased. To overcome it, we utilize Alg. 2 for gray border correction and extra pages elimination in scanned Myanmar text files. Fig. 8 are shown the result of our method for extra page elimination.

Algorithm 2

1. *Read Myanmar Text Image.*
2. *if RGB image then*
   image ⟵ *gray of the input image*
   end.
3. *mask* ⟵ *fill holes of the image.*
4. *for all regions do*
   mask ⟵ *the image of mask is eroded*
   end.
5. *Bs* ⟵ *the regionprops of BoundingBox of Area of image of mask.*
6. *B* ⟵ *the first portion of the Bs.*
7. *for all first to the size of Bs do*
   if the Area of B less than the i^{th} Area of the Bs then
   B ⟵ *the i^{th} of Bs*
   end
   end.
8. *B* ⟵ *the BoundingBox of B.*
9. *image* ⟵ *fit all image pixels into the B.*

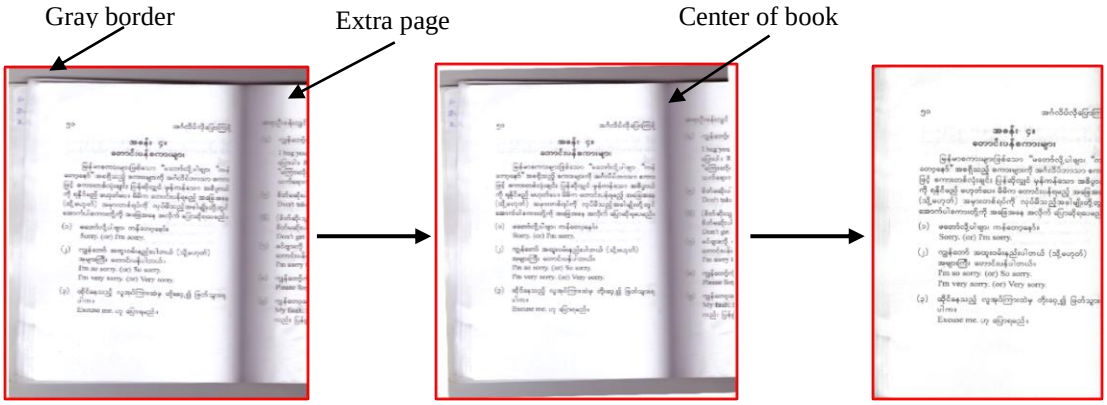


Figure 8. The technique for removal of borders and extra page; (a) Input image with skew and border noises, (b) Free skew image, and (c) Discarding of borders and extra page.

4.6 CHECKING SYSTEM

After mentioning the above tasks, we need to check our system, correct or not. In this portion, our checking system is called line segmentation, an essential step in any OCR system implementation. It involves the line-by-line extraction of all available text in the input image. In [36], the horizontal and vertical projection was used to segment text in input images into lines, words, and characters. Nevertheless, this technique is not very efficient for character extraction in languages like the Myanmar, Gurmukhi, and Devanagari scripts. To this end, the multi-global projection profile technique is adapted for line segmentation in this paper. Multiple lines of text in the same image are split into individual lines using horizontal projection. Alg. 3 is a description of the steps taken to accomplish line segmentation in this paper.

Algorithm 3

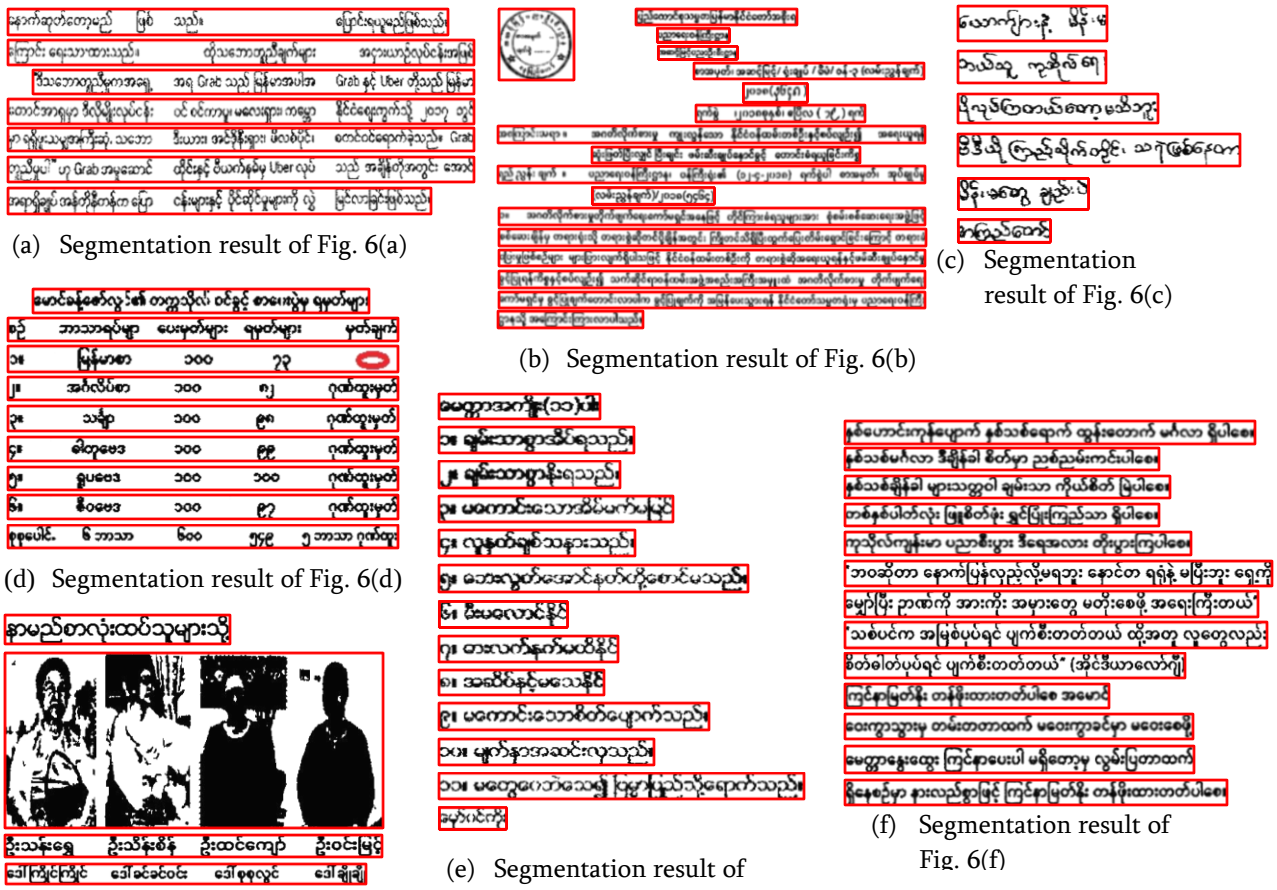
1. *Read Myanmar Text Document.*
2. *orig* ⟵ *resize the input image.*
3. *if RGB then*
   orig ⟵ *the gray of the orig image*
   end.
4. *orig* ⟵ *converts orig image into a binary image.*
5. *orig* ⟵ *discards all small-connected components.*
6. *res* ⟵ *saves change the name of orig.*
7. *i* ⟵ *the first time.*
8. *while 1*
   [fl, res] ⟵ *splits the lines of text in image*
   img ⟵ *fl*
   imag ⟵ *convert img into black on white*
   saves the imag
   if res has finished do
   break
   end
   i ⟵ *i+1*

end.

As an extension method, when input images contain stamps, logos, figures or graphs; our segmentation technique is performed as described in Alg. 4. Though the technique is simple compared to other proposed [37, 38], it gives an excellent result with 100% accuracy. Fig. 9 demonstrates the results of line segmentation on images contain logos, tables, stamps, etc.

Algorithm 4

1. *Read Myanmar Text Document.*
2. *orig ⟵ resizes the input image.*
3. *orig ⟵ converts orig image into a binary image.*
4. *img ⟵ discards all smaller connected components of orig.*
5. *SE ⟵ creates a flat structuring element.*
6. *img1 ⟵ dilates the img1 image with respect to SE.*
7. *[L, N] ⟵ labels and counts the img1 image.*
8. *for i = 1 to N do*
   [r, c] ⟵ finds indexes L, which equal to i in a matrix
   img2 ⟵ sprits texts and photo from an orig image.
   Img3 ⟵ creates an array of all ones for the size of img2 image
   img4 ⟵ subtracts img2 from img3 image.
   img5 ⟵ saves change img4 into black on white image
   saves the img5 image
   end.



(a) Segmentation result of Fig. 6(a)

(b) Segmentation result of Fig. 6(b)

(c) Segmentation result of Fig. 6(c)

(d) Segmentation result of Fig. 6(d)

(e) Segmentation result of Fig. 6(e)

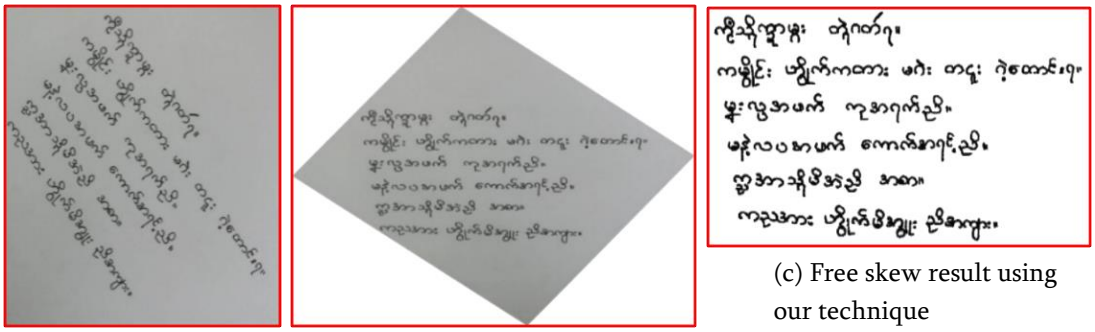
(f) Segmentation result of Fig. 6(f)

(g) Segmentation result of Fig. 6(g)

Figure 9. Sample of line extraction results of Fig. 6.

5. RESULTS AND DISCUSSION

In our experiments, 430 text images were acquired from newspapers, publicly accessible web platforms, books, etc. Texts in these files were written either WinInnwa (WIW), Zawgyi-One (ZG) or Myanmar Unicode (Myanmar Text) fonts and size with 10, 13, 14, 16 and 28. This decision is backed by the knowledge that before 2007; the WinInnwa font type with size 16 was the official font type of Myanmar scripts. Therefore, many important text files, novels, and tabloids contained only this font type and size. From 2007, the Zawgyi-One font type with size 14 became famous and emerged as the most used font type in documents prepared in Myanmar. In 2016, the Myanmar Text (MT) font type became the official font in Myanmar with font size 13. Presently, most web platforms and published files are written both ZG and MT.



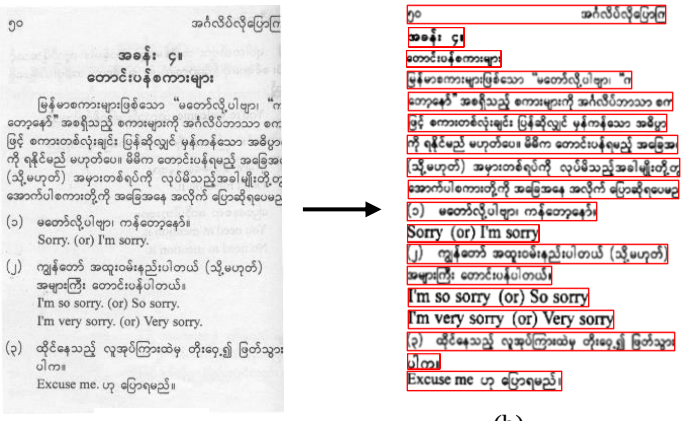
(a) Handwritten input skew image of Mon language

(b) Free skew result using [2, 6, 18, 39] methods

(c) Free skew result using our technique

Figure 10. Comparison of other published methods and ours in free skew.

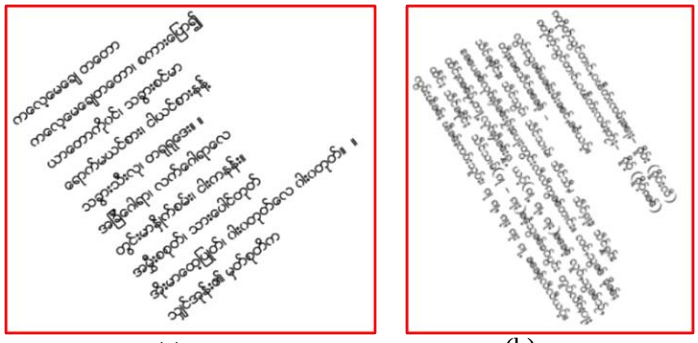
For skew angle detection and free skew, we used various input images as shown in Fig. 5, Fig. 7(a), Fig. 8(a), Fig. 10(a) and Fig. 12. After utilizing Alg. 1 for it, its results discriminate in Fig. 6, Fig. 7(c), Fig. 8(b), and 10(c). To remove borders and extra pages, we used Alg. 2, and its result is in Fig. 8. To check our algorithm in skew angle detection and free skew, we employed Alg. 3 and Alg. 4; their results are shown in Fig. 9, 11 and Fig. 13.



(a) (b)

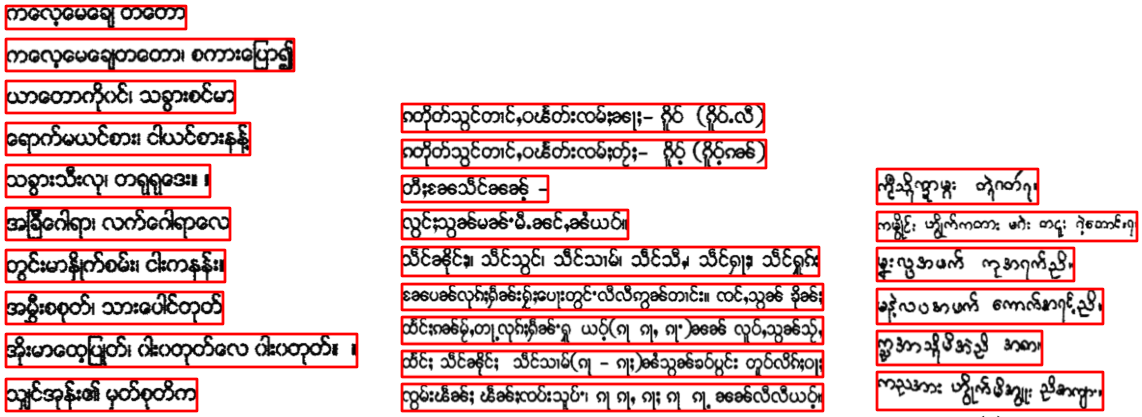
Figure 11. Segmented lines on scanned images from the thick book; (a) Input image from resulted of Fig. 8, and (b) Lines segmentation results of Fig. 11(a).

Accuracy of our technique is enhanced by iterating stages in the algorithms. Though the similar submissions in [2, 6, 18, 39] were capable of finding a skew angle and performing free skew, their results will fail in line segmentation due to unhandled background noise like Fig. 10 in their methods. However, as shown in Fig. 10, our technique properly handled background noise and produced high accuracy result and its line segmentation as in Fig. 13(c). In Fig. 11, we show the result of line segmented image having borders and extra page.



(a) (b)

Figure 12. Various languages with various skew; (a) Rakhine language, and (b) Shan language.



(a) (b) (c)

Figure 13. Results of line extraction technique on various languages; (a) Segmentation results of Fig. 12(a), (b) Segmentation results of Fig. 12(b), and (c) Segmentation results of Fig. 10(c).

Moreover, Table 1 presents the number of images used in this study and the accuracy of our method after iteration. In Table 2, we compare our method with other proposed methods in the literature survey for skew angle correction and checking system. The results show that our method is the most robust and satisfactorily handles skew angle detection and correction with various member of Myanmar languages both printed and handwritten images as well as in the real world.

Table 1. Summary of usable images and accuracy in our methodology.

No.	Types	No. doc.	No. lines	Skew angle	1st free skew	1st checking system	2nd free skew	2nd checking system
1	WIW font with various sizes	30	300	R.S	100	100	100	100
2	ZG font with various sizes	30	300	R.S	100	100	100	100
3	MT font with various sizes	30	300	R.S	100	100	100	100
4	Other languages	50	450	R.S	100	100	100	100
5	Images with multiple column	50	500	R.S	89.1	90.3	100	100
6	Handwritten text images	50	320	R.S	100	100	100	100
7	Images with tables	20	150	R.S	92.2	94.2	99.1	100
8	Images with logos	50	230	R.S	81.2	86.5	100	100
9	Images with multiple skew	20	200	R.S	70.5	72.3	100	100
10	Images with borders and extra page	50	500	-	-	100	100	100
11	Images with background effects	50	300	R.S	100	100	100	100
Total		430	3550	-	93.3%	94.85%	99.92%	100%

Table 2. Comparison of other published methods and ours.

No.	Technique	Testing on printed or handwritten	Images with photos or table	Single or multiple skews	Skew angle limitation	Accuracy	System checking
1	Method [3]	Printed	**Works**	Single	-42º to +42º	**99.9%**	De-check
2	Method [4]	Printed	**Works**	Single	-90º to +90º	**99%**	De-check
3	Method [5]	Printed	**Works**	Single	-45º to +45º	**99.8%**	De-check
4	Method [6]	Printed	**Works**	Single	**No Limitation**	95%	De-check
5	Method [7]	**Both**	Fails	Single	Not Reveal	91%	De-check
6	Method [8]	**Both**	**Works**	Single	-15º to +15º	93.2%	De-check
7	Method [15]	**Both**	**Works**	**Multiple**	Not Reveal	**99.13%**	**99%**
8	Method [39]	Handwritten	Fails	Single	-36º to +36º	92.88%	**97%**
9	Proposed Method	**Both**	**Works**	**Multiple**	**No Limitation**	**99.92%**	**100%**

6. CONCLUSION

We concluded that we firstly proposed a new image purification technique for Myanmar OCR by applying skew angle detection and free skew with high accuracy. In this study, we used 430 images having multiple columns, background photos, stamp, and table within a document alongside the text. Most of the images used were written in Myanmar text, but other languages like Chinese and Russian and so on can use our technique. Our skew angle detection and free skew implementation give an almost 100% accuracy. In addition, we got 100% accuracy for lines extraction on the Myanmar text images. Images with multiple skews and no skew are also properly handled by our technique. With our present results, our work is can be enhanced towards achieving an efficient Myanmar OCR system as well as other languages. We aim to present an efficient Myanmar OCR system capable of processing scanned documents with different font types, font sizes, tables, background photos, stamps, global or multiple skew angles, and various background noises.

Our method, skew angle detection and free skew will be down-and-out accuracy on images having curvilinear text lines. In addition, some characters, touching with center can lose information when the input images are collected from thick books. Moreover, if tables in images include mathematical signs like +, – and ×, our algorithm cannot recognize it. In addition, if the input images have only photos without text lines, our checking system does not work.

REFERENCES

[1] T. Jundale, R. Hegadi, Research survey on skew detection of Devanagari script, International Journal of Computer Applications, National Conference on Knowledge, Innovation in Technology and Engineering (NCKITE), 2015, 41-44.

[2] M. Basavanna, S. S. Gornale, Skew detection and skew correction in scanned document image using principal component analysis, International Journal of Scientific & Engineering Research (IJSER), Vol. 6, Issue 1, 2015, 1414-1417.

[3] A. Papandreou, B. Gatos, S. J. Perantonis, I. Gerardis, Efficient skew detection of printed document images based on novel combination of enhanced profiles, IJDAR 17, Springer, 2014, 433-454.

[4] N. Watts, J. Rani, Performance evaluation of improved skew detection and correction using FFT and Median filtering, International Journal of Computer Applications (IJCA), Vol. 100, No. 15, 2014, 7- 16.

[5] O. Boudraa, W. K. Hidouci, D. Michelucci, An improved skew angle detection and correction technique for historical scanned documents using morphological skeleton and progressive probabilistic Hough Transform, 5th International Conference on Electrical Engineering-Boumerdes (ICEE-B), IEEE, 2017, 1-6.

[6] M. Shafii, M. Sid-Ahmed, Skew detection and correction based on an axes-parallel bounding box, IJDAR, Vol. 18, Springer, 2014, 59-71.

[7] T. A. Jundale, R. S. Hegadi, Skew detection of Devanagari script using pixels of the axes-parallel rectangle and linear regression, International Conference on Energy Systems and Application (ICESA), IEEE, 2015, 480-484.

[8] A. Alaei, P. Nagabhushan, U. Pal, F. Kimura, An efficient skew estimation technique for scanned documents: an application of piece-wise painting algorithm, JOURNAL OF PATTERN RECOGNI-TION RESEARCH 1, 2016, 1-14.

[9] C. S. Lwin, X. Wu, Zone-wise segmentation and lexicon-driven recognition for printed Myanmar characters, International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), Vol. 3, No. 8, 2018, 161-180.

[10] S. N. Holambe, Dr. U. B. Shinde, S. D. Mali, Reorganization of Devanagari script character using genetic algorithm, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Vol. 6, Issue 5, 2017, 736-743.

[11] P. Nehete, A survey on estimation & correction of multiple skews in document image processing, International Journal of Current Trends in Engineering & Research (IJCTER), Vol. 2, Issue 3, 2016, 103-106.

[12] A. Papandreou, B. Gatos, A novel skew detection technique based on vertical projections, International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2011, 384-388.

[13] H. P. P. Win, K. N. N. Tun, Converting Myanmar printed document image into machine understandable text format, 6th International Conference on Digital Information Management, IEEE, 2011, 96-101.

[14] S. W. Mohammed, N. R. Soora, Global skew detection and correction using morphological and statistical methods, Computational Vision and Bio Inspired Computing, Springer, 2018, 556-568.

[15] N. R. Soora, P. S. Deshpande, A novel local skew correction and segmentation approach for printed multilingual Indian documents, Alexandria Engineering Journal, Vol. 57, Issue 3, 2018, 1609-1618.

[16] R. Singh, R. Kaur, Improved skew detection and correction approach using discrete Fourier algorithm, International Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231-2307, Vol. 3, Issue 4, 2013, 5-7.

[17] A. Boukharouba, A new algorithm for skew correction and baseline detection based on the randomized Hough Transform, Journal of King Saud University Computer and Information Sciences, Production and hosting by Elsevier B. B., Vol. 29, Issue 1, 2017, 29-38.

[18] K. C. Prakash, Y. M. Srikar, G. Trishal, S. Mandal, S. S. Channappayya, Optical character recognition (OCR) for Telugu: database, algorithm, and application, 25th IEEE International Conference on Image Processing (ICIP), 2018, 3963-3967.

[19] F. Md. Hasan, T. Afroz, S. Ismail, S. Md. Islam, Document decomposition of Bangla printed text, 4th International Conference on Engineering Research, Innovation and Education (ICERIE), 2017.

[20] A. AL-Khatatneh, S. A. Pitchay, M. AI-qudah, A review of skew detection techniques for document, 17th UKSIM-AMSS International Conference on Modelling and Simulation, IEEE, 2015, 316-321.

[21] B. Jain, M. Borah, A survey paper on skew detection of offline handwritten character recognition system, International Journal of Computer Engineering and Applications, Vol. VI, Issue I, 2014.

[22] R. N. Verma, Dr. L. G. Malik, Review of illumination and skew correction techniques for scanned documents, Procedia Computer Science, Vol. 45, 2015, 322-327.

[23] http://www.worldometers.info/world-population/myanmar-population/.

[24] Goddard, Cliff, The languages of East and Southeast Asia: An introduction, Oxford University Press, ISBN 0-19-924860-5, 2005.

[25] https://en.wikipedia.org/wiki/Burmese_language.

[26] https://www.unicode.org/charts/PDF/U1000.pdf.

[27] D. Brodic, C. A. B. Mello, C. A. Maluckov and Z. N. Milivojevic, An approach to skew detection of printed documents, Journal of Universal Computer Science (J.UCS), Vol. 20, No. 4, 2014, 488-506.

[28] N. Otsu, A threshold selection method from gray-level histograms, IEEE Transactions on Systems, Man, and Cybernetics, Vol. 9, No. 1, 1979, 62-66.

[29] https://ww2.mathworks.cn/help/images/ref/imgaussfilt.html#bunfgk6-1-sigma.

[30] R. A. Haddad, A. N. Akansu, A class of fast Gaussian binomial filters for speech and image

processing, Transactions on Acoustics, Speech and Signal Processing, vol. 39, March 1991, 723-727.

[31] https://en.wikipedia.org/wiki/Gaussian_filter #cite_note-NixonAguado-6.

[32] K. Arulmozhi, S. A. Perumal, C. S. T. Priyadarsini, K. Nallaperumal, Image refinement using skew angle detection and correction for Indian license plates, International Conference on Computational Intelligence and Computing Research, IEEE, 2012.

[33] https://en.wikipedia.org/wiki/Linear_equation, 2018.

[34] https://en.wikipedia.org/wiki/Slope, 2018.

[35] https://en.wikipedia.org/wiki/Hough_transform #cite_note-9, 2018.

[36] D. A. Noola, M. M. Kodabagi, An approach to extract line, word and character from scene text image, International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE), Vol. 14, Issue 2, 2015, 916-922.

[37] V. Yadav, N. Ragot, Text extraction in document images: highlight on using corner points, 12th IAPR Workshop on Document Analysis Systems, IEEE, 2016, 281-286.

[38] J. Zhang, Y. Zhu, J. Du, L. Dai, Trajectory-based radical analysis network for online handwritten Chinese character recognition, 24th International Conference on Pattern Recognition (ICPAR), IEEE, 2018, 3681-3686.

[39] T. A. Jundale, R. S. Hegadi, Skew detection and correction of Devanagari script using Hough Transform, Procedia Computer Science, Vol. 45, 2015, 305-311.

Authors

Chit San Lwin holds a BSc degree from Monywa University, Myanmar, from 2006. Received MSc from Moscow State University (MSU), Russia, in 2011. Graduated with a master of research (MRes) from Monywa University, Myanmar, in 2015. Currently pursuing a Ph.D. degree at Harbin Institute of Technology (HIT), Harbin, China.

Wu Xiangqian has been a professor at School of Computer Science and Technology, Harbin Institute of Technology, China, since 2009. He has authored one book and over 100 papers in international journals and conferences. Current research interests include computer vision, pattern recognition, and biometrics and medical image analysis. He is also my supervisor.

Cite this article as:

Chit San Lwin, Wu Xiangqian, "Image Purification Technique for Myanmar OCR Applying Skew Angle Detection and Free Skew", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN: 2395-602X, Print ISSN: 2395-6011, Volume 6 Issue 1, pp. 186-203, January-February 2019.

Available at doi :

https://doi.org/10.32628/IJSRST19615
Journal URL : http://ijsrst.com/IJSRST19615